DOCUMENT RESUME

ED 048 357 TM 000 433

AUTHOR Keats, John B.; Brewer, James K.
TITLE A Distribution-Free Test for Model Comparisons.
PUB DATE Feb 71
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association, New York, New York, February 4-7, 1971

EDRS PRICE EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS *Goodness of Fit, Hypothesis Testing, *Mathematical Models, *Nonparametric Statistics, Probability Theory, Research Methodology, *Statistical Analysis, *Tests of Significance

ABSTRACT

This paper presents an index of goodness-of-fit for comparing m models over n trials. The index allows for differentiated weighting of the trials as to their importance in the comparison of the models. Several possible weighting schemes are suggested and the conditions on the weights which assure asymptotic normality of the index distribution are presented. A relative goodness-of-fit test using the index is proposed which is distribution-free under the null hypothesis. Both single model and simultaneous inferential tests are presented for large values of n. An application of the index and subsequent inferences are provided using three probability learning models and human subject data. (Author)

A Distribution-Free Test for Model Comparisons

by

John B. Keats
Louisiana Tech University
Ruston, Louisiana

and

James K. Brewer
Florida State University
Tallahassee, Florida

Paper presented at the Annual Meeting
of the American Educational Research
Association, New York, N.Y.

February 5, 1971

A Distribution-Free Test for Model Comparisons

1. Introduction

Consider a finite set of m mathematical models which have each provided estimates of subject data at n trial points. A general problem of model comparisons is concerned with deciding if any one model is a better "fit" of the data than the other models when there is no universally accepted yardstick of "fit" or standard statistical test [Bush and Mosteller, 1959]. The most common approach to the problem is to compare each model to the data using some χ^2-like procedure. The basic assumption involved therein is one of independence across trials (or blocks of trials) in order to satisfy the additivity of the statistical test model. However, if the models are in any way path-dependent and it is expected that the fit of the models are functions of n, then the use of such comparison tests is inappropriate since some trials would be more important than others for model comparison purposes. Tests of the Kolmogorov-Smirnov, Cramér-von Mises type (e.g., Birnbaum, 1953; Darling, 1957; Massey, 1951) and others (e.g., Anderson & Darling, 1954; Riedwyl, 1967; Tsao, 1955) require a continuous cumulative distribution function for the random variable which accounts for the data. Atkinson (1969) presents several tests for model comparisons which measure the deviations of each model's predictions from some "best" formula which is found by regression.

This paper proposes a distribution-free index for model comparisons which makes no assumptions about continuity. The index also allows for differential weighting of trials according to the effect of each trial on the data. For example, suppose one is comparing several learning models and it can be assumed

that each model's proximity (as measured by some definition of proximity) to the data is a monotonically increasing function of n. Then the trials over which the models are compared should be differentially weighted by giving heavier weights to the later trials; for it is in these trials that the models are expected to provide better estimates of the data points.

Let the measure of closeness of model j at trial i be defined as

$$f_{ij} = \left| \hat{y}_{ij} - y_i \right|, \quad i = 1, 2, \ldots, n; \quad j = 1, 2, \ldots, m,$$

where $\hat{y}_{ij}$ is the estimate of the data value y_i given by model j at trial i. It is apparent that if model j is a closer fit of the data than the other m-1 models, then f_{ij} values will generally be smaller than f_{ik} values, $k \neq j$ (and vice versa for a poorer fit).

The goodness-of-fit index, defined in Section 2, assigns positive integer ranks, r_{ij}, to the f_{ij} for all i and j. Then weights, w_i, are assigned to the comparison points based on theoretical or empirical considerations as to the importance of each trial for comparison purposes. Some properties, including the conditions on the w_i for asymptotic normality of the index distribution are given. In Section 3, a discussion of large sample single-model and simultaneous inference is presented. Section 4 suggests several possible permissible weighting schemes and Section 5 illustrates the procedure with three probability learning models.

2. The Index and Its Properties

Let the index, denoted, I_j, be defined for the jth model as

$$I_j = \sum_{i=1}^{n} \frac{(m - r_{ij})\, P_i}{m - 1} \quad , \tag{1}$$

where m = number of models under consideration,

r_{ij} = rank of f_{ij} for the jth model at trial i,

$$P_i = \frac{w_i}{\sum_{k=1}^{n} w_k} \quad \text{and}$$

w_i = weight assigned to trial i.

Some Properties of I_j:

Proofs of these properties have been ommitted for the sake of brevity.

(2) a) $0 \leq I_j \leq 1$.

b) If $r_{ij} = k$ for all i then,

(3) $$I_j = \frac{m - k}{m - 1}$$

c) The maximum non-perfect ($I_j \neq 1$) value of I_j will occur when model j is ranked 2 for the data point having the smallest weight and ranked 1 for all other data points. If rank w_t is given the tth data point then the maximum non-perfect I_j value is given be

(4) $$1 - \frac{w_t}{(m - 1)\sum_{i=1}^{n} w_i}$$

(5) d) $$\sum_{j=1}^{m} I_j = \frac{m}{2}$$

e) If it is assumed that model j fits the data no better than the other models over all trials, then

(6) $$E(I_j) = 1/2 \ .$$

f) If the rank value of model j at trial i is independent of the rank value of model j at trial i+k ($k > 0$) and model j fits the data no better than any other model, i.e., $P(r_{ij}) = 1/m$ for all i, then

(7) $$VAR(I_j) = \frac{\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2} \left[\frac{m+1}{12(m-1)}\right]$$

g) A direct application of the Lindenberg-Feller Theorem [Gnedenko and Kolmogorov, 1954] shows that if

(8) $$\lim_{n\to\infty} \frac{\max w_i}{\sqrt{\sum_{i=1}^{n} w_i^2}} = 0 \ ,$$

then I_j is asymptotically normally distributed. The converse can also be shown to hold.

3. Significance Tests with I_j

The implications of property g) is Section 2 is that, coupled with the independence assumption of property f), we can, for sizable n, use I_j to conduct a test of

(9) $$H_o: \ P(f_{ij} > f_{ik}) = P(f_{ij} < f_{ik}), \quad j \neq k, \ j \text{ fixed},$$

which is distribution-free under H_o. Note that this H_o is equivalent to

H_o: $P(r_{ij}) = P(r_{ik})$ for all i, $j \neq k$.

Since, under H_o, I_j is asymptotically normally distributed with mean and variance given respectively by (6) and (7), the statistic

$$z_{I_j} = \frac{I_j - E(I_j)}{\sqrt{VAR(I_j)}}$$

is approximately distributed N(0,1).

When confronted with more than one hypothesis, for example, when j is not fixed in (9), a device for scaling down the significance level can be used. One such device, resulting from the Bonferroni Inequality [Miller, 1966], suggests the $\alpha/2m$ level of significance for simultaneous two-tailed tests. Crude though this estimated significance level is, its derivation does not depend on the I_j, j = 1, 2, ..., m, being independent as do most simultaneous test approximations.

4. Some Weighting Functions

The basic subjective portion in the development and use of index I_j is the assignment of weights w_i, i = 1, 2, ..., n to the trials. This will depend on the relative importance which the experimenter places on the trials used for the comparisons of the models and can take on almost any functional form. There are, however, several which (a) satisfy the condition in expression (8) for asymptotic normality, (b) are rational, and (c) possess mathematical simplicity. Three such weighting functions and their resulting variances are herein presented.

Function 1: $w_i = c$, $c \neq o$.

The effect here is one of proposing that the trials are all equal for comparison purposes. This would be the case if an experimeter assumed random

behavior models and did not expect the models to be better fits toward the end of the trials than at the beginning of the trials or vice versa.

Here the variance of I_j under the assumptions of property f) becomes

$$\text{(10)} \qquad \text{VAR}(I_j) = \frac{m + 1}{12n(m - 1)}$$

Function 2: $W_i = i$

For this case the assumption is that the later comparison trials are more important than the earlier trials. This would be appropriate if an experimenter felt that the models required the earlier trials to sequentially reach a point beyond which the comparisons with data would be reasonable.

Under this scheme and the assumptions of property f),

$$\text{(11)} \qquad \text{VAR}(I) = \frac{(2n + 1)(m + 1)}{18(n + 1)(n)(m - 1)}$$

Function 3: $W_i = n - i + 1$

The assumption for this scheme is that the earlier trials are more important for model comparison purposes than the later trials and would be appropriate to use if one believed that some kind of "fatigue" factor was involved. For example, suppose it was suspected that beyond trial k, the behavior being modeled gradually began to act in a random or erratic fashion. Then the later trials could be thought of as "unreliable" for model comparisons. The variance here is the same as under Function 2 except for subsets of trials in which the summation of (1) does not run over the full range from 1 to n.

5. Example

Three probability learning models were compared over the last 10 trials

of a two-choice experiment. In this experiment, human subjects were asked to predict which of two possible events, E_1 or E_2, would occur on each of a series of trials. Predictions of E_1 and E_2 are denoted by responses A_1 and A_2, respectively. At the end of each trial, the subjects were permitted to observe which event actually occurred. Event E_1 had a fixed probability $\pi = .7$ of occurring in a random sequence.

Model 1, a linear operator model, is due to Estes (1950) and in this experiment assumed the form

$$\text{(12)} \qquad \Pr(A_1 \text{ on trial } n+1) = P_{1,n+1} = \pi - [\pi - P_{1,n}](1 - \theta)^{n-1}$$

where θ is a rate of learning parameter, $0 \leq \theta \leq 1$, estimated from observed response frequencies. This model is usually applied to experiments with many more trials than the experiment of this example, but is included here for illustrative purposes only.

Models 2 and 3 are of the form

$$\text{(13)} \qquad P_{1,n+1} = e \cdot \mu'$$

where e is an experience vector of length n representing the trials 1, 2, ..., n and composed of the digits one or zero depending on whether or not E_1 occurred on a particular trial, and μ is a memory vector whose n elements are proportional to the probabilities of recalling the events of trials 1 to n such that $\sum_{i=1}^{n} \mu_i = 1$. The development of the basic theory relative to (13) is due to Overall (1960).

Models 2 and 3 differ in the estimation of elements in the memory vector, μ. Model 2 employed probabilities which were empirically determined by Murdock (1962) under a variety of recall conditions, none of which involved

binary items. The probability of recalling the ith item in a sequence of length n was

$$(14) \qquad P(i,n) = 1.00 + .27e^{-.77(i-1)} - .772(.042)^{.555\,(n-i)} \quad .$$

The probabilities used in Model 3 were estimated from the data of a previous experiment where subjects engaged in the two-choice task made recalls at periodic intervals. These probabilities were

$$(15) \qquad P(i,n) = .9047 - .0694(i) + .0775(i^2/n) - .3577(i/n)^2 \quad .$$

Thirty-eight subjects performed the two-choice experiment with $\pi = .7$. Table 1 presents the results of the last 10 trials.

Suppose that Model 3 is of particular interest. If a test of

$$H_o: \; P(f_{i3} > f_{ik}) = P(f_{i3} < f_{ik}) \;, \quad i = 1, 2, \ldots, 10 \;, \quad k = 1,2 \;,$$

is conducted at $\alpha = .05$ under the weights of Function 2, $I_3 = .8000$ and $Z_{I_3} = \frac{.8 - .5}{.1457} = 2.059$. Since $P(z \geq 2.059) \cong .02$, H_o is rejected and Model 3 can be judged to be a significantly better fit than the other models. However, with the weights of Function 1, $I_3 = .7500$ and $Z_{I_3} = \frac{.75 - .50}{.4083} = .637$, so that $P(z \geq .637) \cong .26$. Thus H_o cannot be rejected. Models 1 and 2 are not significant under either weighting function since their index values are .4545 and .2455 respectively using Function 2, and .45 and .30 respectively with Function 1. The Bonferroni test was not significant at the .05 level using Functions 1 or 2.

TABLE 1

COMPARISON OF THREE PROBABILITY LEARNING MODELS

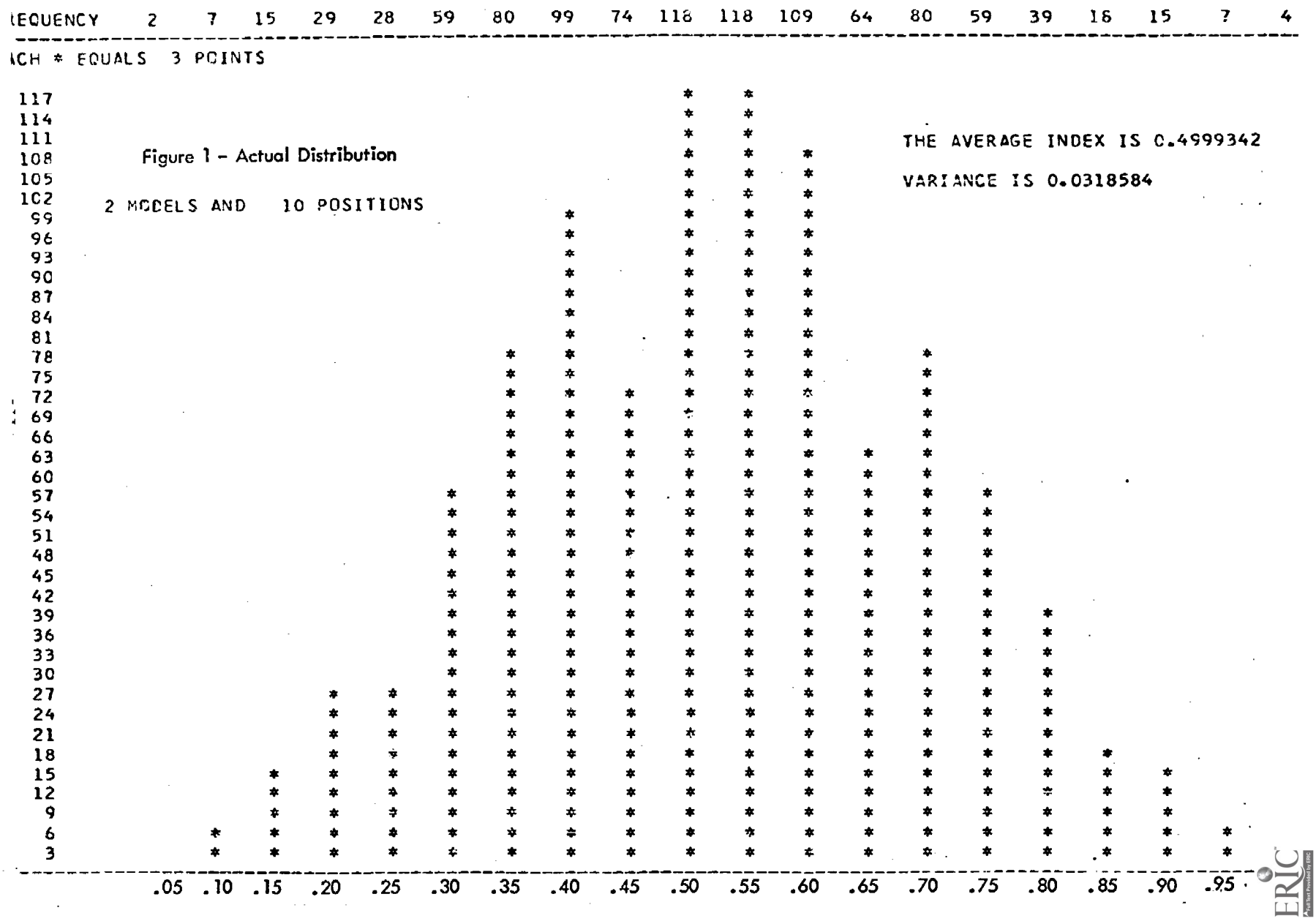
trial (j)	Proportion Predicting E_1									
	Model 1 (θ = .05)	Model 2	Model 3	Actual	f_{1_j}	rank	f_{2_j}	rank	f_{3_j}	rank
1	.5986	.5357	.6135	.7368	.1382	2	.2011	3	.1233	1
2	.6037	.6125	.6473	.5789	.0248	1	.0366	2	.0684	3
3	.6086	.6880	.6761	.6842	.9756	3	.0038	1	.0081	2
4	.6132	.7493	.7008	.6842	.0710	3	.0651	2	.0166	1
5	.6176	.7924	.7220	.7368	.1192	3	.0556	2	.0148	1
6	.6218	.8208	.7403	.6842	.0624	2	.1366	3	.0561	1
7	.6258	.8392	.7562	.5263	.0995	1	.3129	3	.2299	2
8	.6296	.7138	.7059	.5263	.1033	1	.1875	3	.1796	2
9	.6332	.7405	.7240	.6842	.0510	2	.0563	3	.0398	1
10	.6366	.6423	.6795	.7894	.1528	3	.1471	2	.1099	1

Comments and Discussion

In any procedure involving assigned ranks the problem of tied rankings merits discussion. We have purposely avoided the issue since the standard procedures of randomly breaking ties or assigning average ranks are quite satisfactory for small numbers of ties. If the number of ties is large, say > 20%, then we recommend that I_j not be used as an index of goodness-of-fit. No adjustments are herein proposed to account for the number of ties in calculating I_j.

The tests of significance presented in Section 3 were exclusively for "large" sample sizes, and we propose that $n \geq 10$ is sufficient to warrant the use of the z statistic for testing purposes. Figure 1 graphically illustrates the exact distribution of I_1 under H_o for $n = 10$, $m = 2$, and $w_i = i$, $i = 1, 2, \ldots, 10$. It appears that the normal distribution would yield quite reasonable approximations. A more practical reason for not deriving and providing exact distributions for $n \leq 10$ is purely financial. The reader appreciates the massive computing job necessary to provide these distributions for $m > 2$ and the need for such small sample distributions (since number of trials usually exceeds 10) does not justify the expenditure at present. Only if one were estimating a small number of parameters by several models would exact small-sample distributions be desirable.

The primary concern in this paper has been with the presentation of an index of relative goodness-of-fit rather than its accompanying tests of significance. This was due to the restrictive assumptions underlying the test of H_o. The testable hypothesis itself may be so general in form that


FREQUENCY 2 7 15 29 28 59 80 99 74 118 118 109 64 80 59 39 18 15 7 4
EACH * EQUALS 3 POINTS
117
114
111
108
105
102
99
96
93
90
87
84
81
78
75
72
69
66
63
60
57
54
51
48
45
42
39
36
33
30
27
24
21
18
15
12
9
6
3
Figure 1 – Actual Distribution
2 MODELS AND 10 POSITIONS
THE AVERAGE INDEX IS 0.4999342
VARIANCE IS 0.0318584
.05 .10 .15 .20 .25 .30 .35 .40 .45 .50 .55 .60 .65 .70 .75 .80 .85 .90 .95

a researcher would not want to test it in the first place. This, as in any test of hypothesis, does not invalidate the use of the index with its accompanying properties for descriptive purposes as a goodness-of-fit indicator.

The aspect of selecting weight functions is obviously a crucial one and we have presented only a small selection of simple functions. We have also assumed that the same weight function would be used for each model for a particular comparison. The problem of assigning a priori a "best" function per model and then making the comparisons was not addressed and indeed would be a difficult problem to handle statistically.

REFERENCES

Anderson, T. W. & D. A. Darling, "A test of goodness of fit." Journal of the American Statistical Association, 49 (1954), 765-69.

Atkinson, A. C., "A test for discriminating between models." Biometrika, 56, 2(1969), 337-47.

Birnbaum, Z. W., "Distribution-free tests of fit for continuous distribution functions." Annals of Mathematical Statistics, 24 (1953), 1-8.

Bush, R. R. & F. Mosteller, A comparison of eight models. In R. Bush and W. Estes (eds.), Studies in Mathematical Learning Theory, Stanford: Stanford University Press, 1959.

Darling, D. A., "The Kolmogorov-Smirnov, Cramér-von Mises tests." Annals of Mathematical Statistics, 28 (1957), 823-38.

Estes, W. K., "Toward a statistical theory of learning." Psychological Review, 57 (1950), 94-107.

Massey, F. J., "The Kolmogorov-Smirnov test for goodness of fit." Journal of the American Statistical Association, 46 (1951), 68-78.

Miller, Rupert G., Simultaneous Statistical Inference, New York: McGraw-Hill, 1966.

Murdock, B. B., "The serial position effect of free recall." Journal of Experimental Psychology, 64, 5 (1962), 482-88.

Overall, J. E., "A cognitive probability model for learning." Psychometrika, 25 (1960), 159-172.

Riedwyl, Hans, "Goodness of fit." Journal of the American Statistical Association, 62 (1967), 390-98.

Tsao, C. K., "Rank sum tests of fit." Annals of Mathematical Statistics, 26 (1955), 94-104.

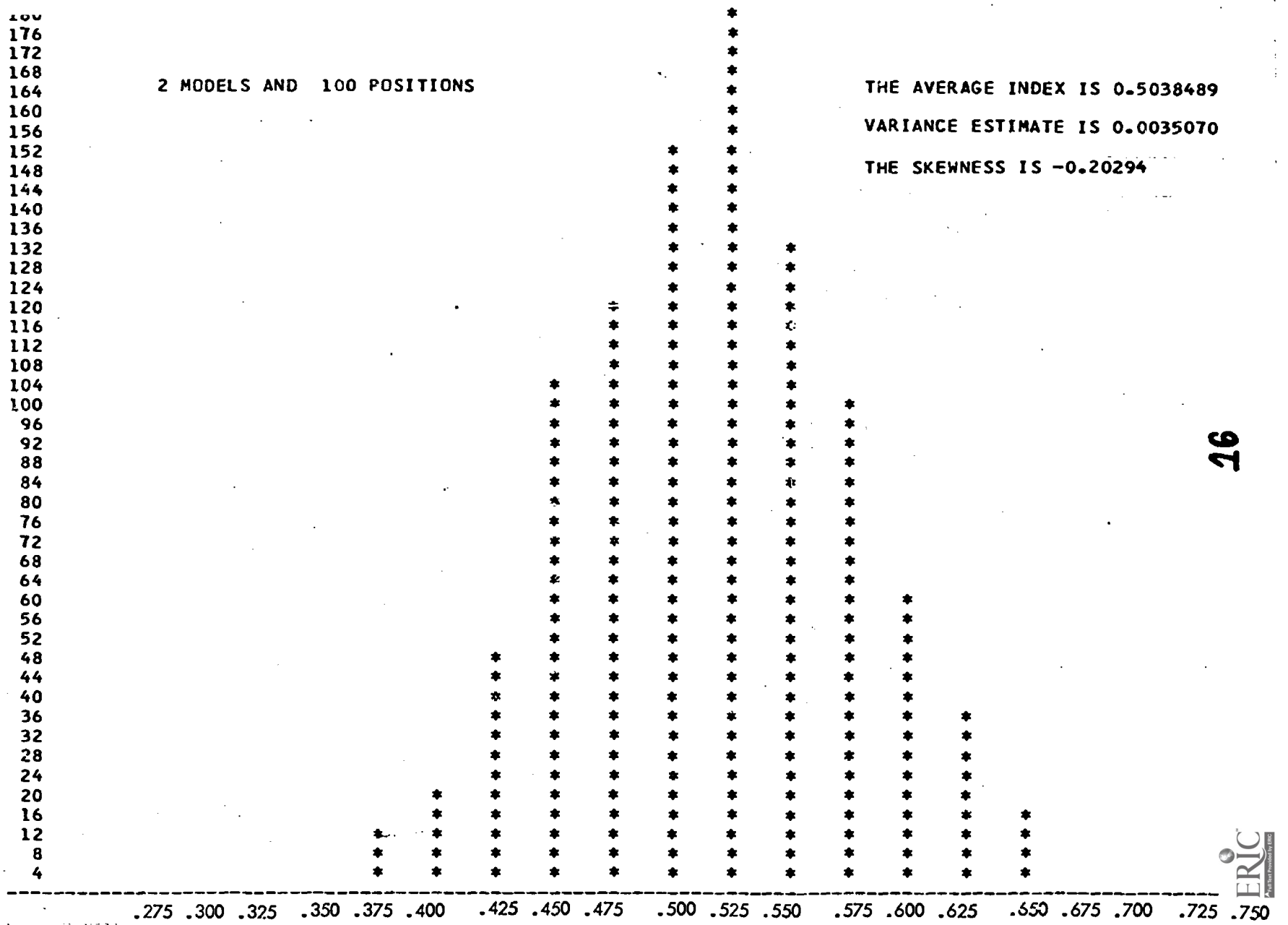
Gnedenko, B. V. & Kolmogorov, A. N. Limit Distributions for Sums of Independent Random Variables, Cambridge, Massachusetts: Addison-Wesley, 1954.

APPENDIX

Simulated Distributions of any I_j for

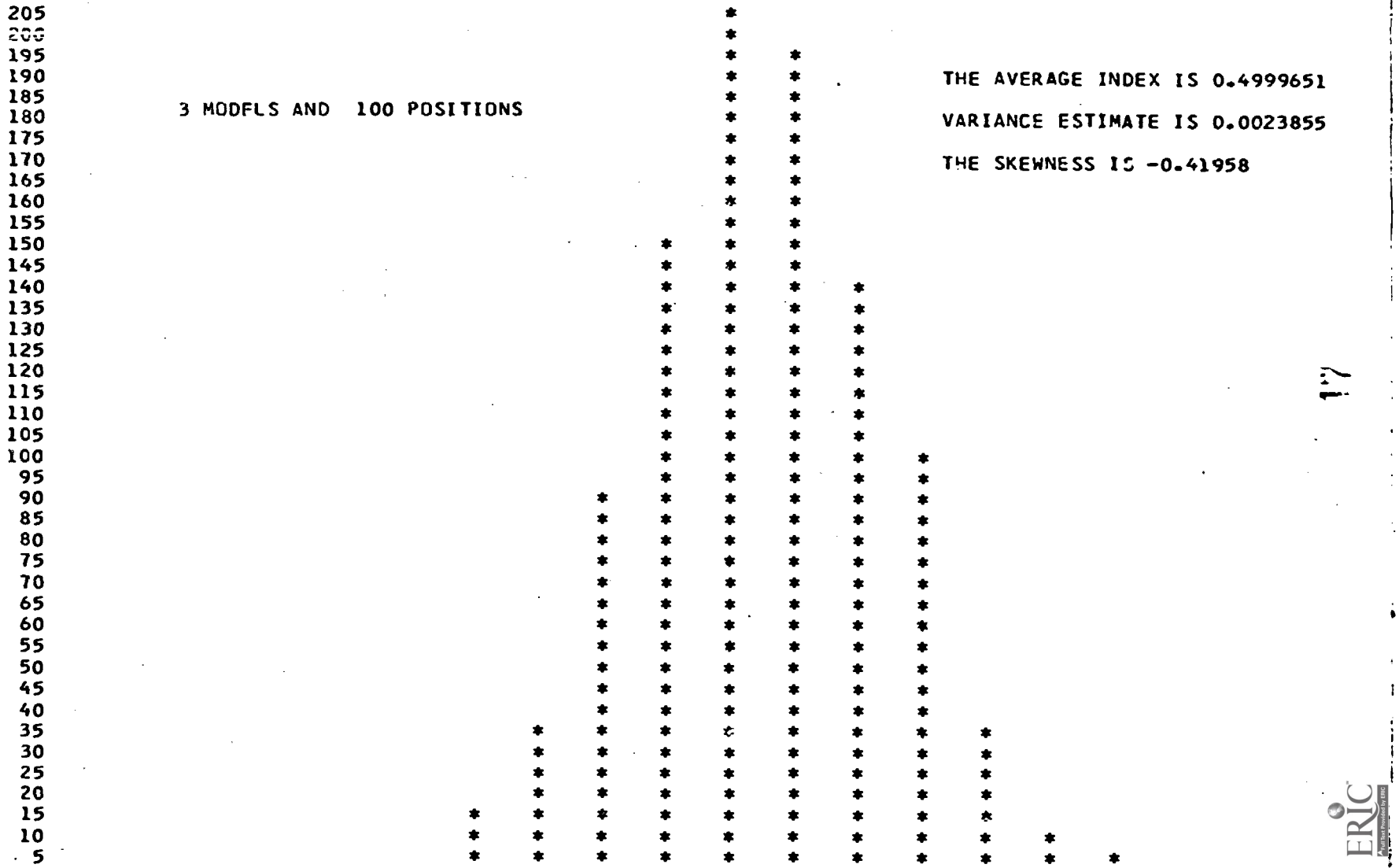
m = 2, 3, 4, 5 over 100 Trials Assuming

$Pr(r_{ij}) = 1/m$ for i = 1, 2, ...100.


2 MODELS AND 100 POSITIONS
THE AVERAGE INDEX IS 0.5038489
VARIANCE ESTIMATE IS 0.0035070
THE SKEWNESS IS -0.20294
176
172
168
164
160
156
152
148
144
140
136
132
128
124
120
116
112
108
104
100
96
92
88
84
80
76
72
68
64
60
56
52
48
44
40
36
32
28
24
20
16
12
8
4
.275 .300 .325 .350 .375 .400 .425 .450 .475 .500 .525 .550 .575 .600 .625 .650 .675 .700 .725 .750

EQUENCY	0	0	0	2	2	15	36	90	154	207	196	141	101	38	13	5	0	0	0	0

CH * EQUALS 5 POINTS

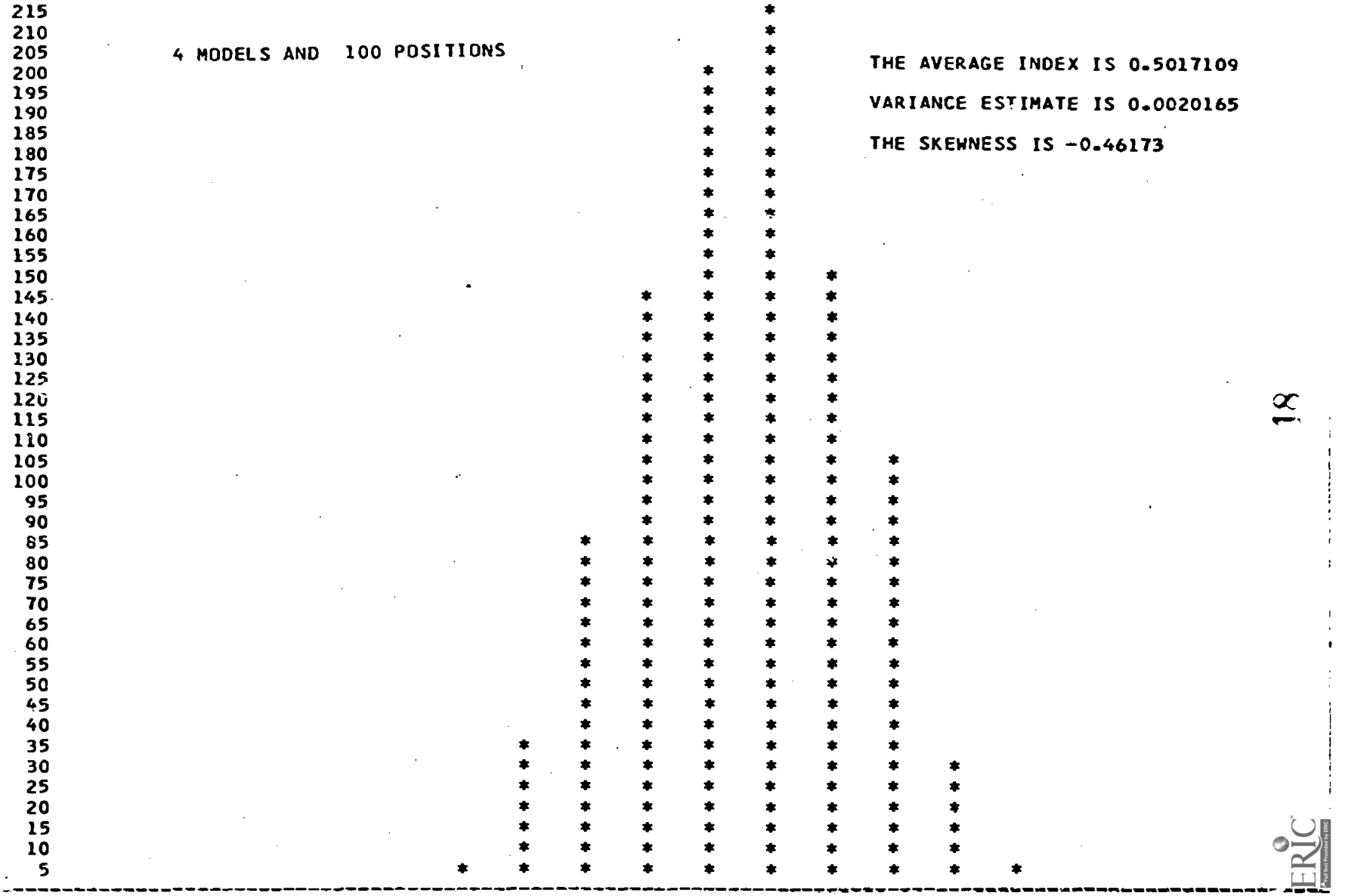


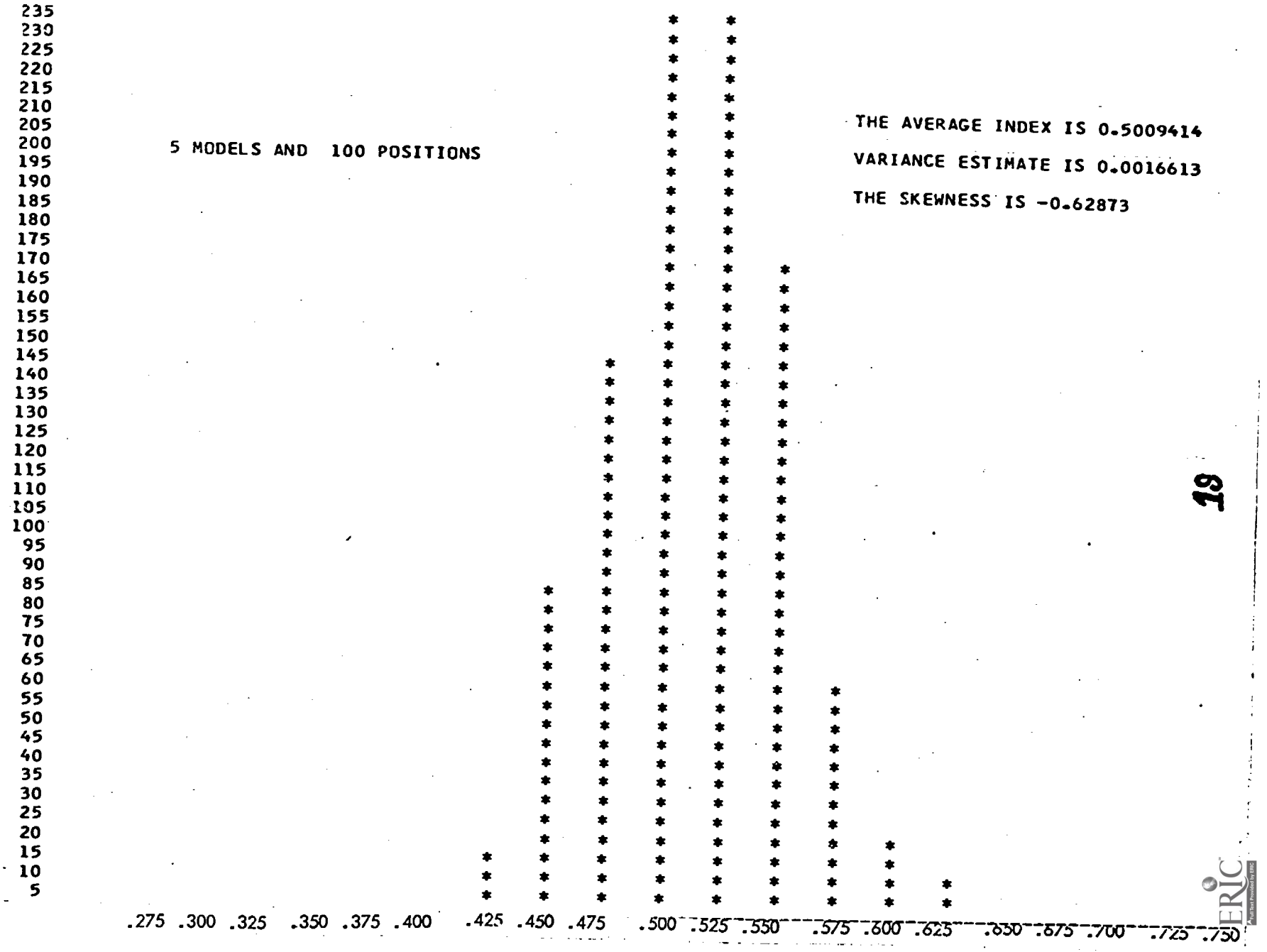
3 MODELS AND 100 POSITIONS

THE AVERAGE INDEX IS 0.4999651

VARIANCE ESTIMATE IS 0.0023855

THE SKEWNESS IS -0.41958


4 MODELS AND 100 POSITIONS
THE AVERAGE INDEX IS 0.5017109
VARIANCE ESTIMATE IS 0.0020165
THE SKEWNESS IS -0.46173
215
210
205
200
195
190
185
180
175
170
165
160
155
150
145
140
135
130
125
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
.275 .300 .325 .350 .375 .400 .425 .450 .475 .500 .525 .550 .575 .600 .625 .650 .675 .700 .725 .750


5 MODELS AND 100 POSITIONS
THE AVERAGE INDEX IS 0.5009414
VARIANCE ESTIMATE IS 0.0016613
THE SKEWNESS IS -0.62873
235
230
225
220
215
210
205
200
195
190
185
180
175
170
165
160
155
150
145
140
135
130
125
120
115
110
105
100
95
90
85
80
75
70
65
60
55
50
45
40
35
30
25
20
15
10
5
.275 .300 .325 .350 .375 .400 .425 .450 .475 .500 .525 .550 .575 .600 .625 .650 .675 .700 .725 .750